The connection between Bayesian estimation of a Gaussian random field and RKHS

Aleksandr Y. Aravkin, Bradley M. Bell, James V. Burke and Gianluigi Pillonetto

***Abstract*—Reconstruction of a function from noisy data is often formulated as a regularized optimization problem over an infinite-dimensional reproducing kernel Hilbert space (RKHS). The solution describes the observed data and also has a small RKHS norm. When the data fit is measured using a quadratic loss, this estimator has a known statistical interpretation in terms of Gaussian random fields: it provides the minimum variance estimate of the unknown function given the noisy measurements. In this paper, we provide a statistical interpretation when more general losses are used, such as Vapnik or Huber. For a given data set, the RKHS function estimate contains all the possible maximum a posteriori estimates of a random field for which the prior distribution is Gaussian.**

***Index Terms*—kernel based regularization; Gaussian processes; representer theorem; reproducing kernel Hilbert spaces; regularization networks; support vector regression**

I. Introduction

Minimizing a regularized functional with respect to a reproducing kernel Hilbert space (RKHS) $\mathscr{H}$ is a popular approach to reconstruct a function $F : \mathscr{X} \to \mathbf{R}$ from noisy data; e.g. see [2], [17], [19], [6]. To be specific, regularization in $\mathscr{H}$ estimates F using $\hat{F}$ defined by

$$\hat{F} = \arg\min_{F \in \mathscr{H}} \left(\sum_{i=1}^{N} V_i[y_i - F(x_i)] + \gamma \|F\|^2_{\mathscr{H}} \right) , \qquad (1)$$

where $\gamma \in \mathbf{R}^+$ is the regularization parameter, $\mathscr{X}$ is a set (possibly finite), $x_i \in \mathscr{X}$ is the location where $y_i \in R$ is measured, $V_i : \mathbf{R} \to \mathbf{R}^+$ is the loss function for y_i, and $\|\cdot\|_{\mathscr{H}}$ is the RKHS norm induced by the positive definite reproducing kernel $K : \mathscr{X} \times \mathscr{X} \to \mathbf{R}$, see [2]. Notice that $x_i \notin \mathbf{R}$ and is not a component of a vector x, as is the convention for other subscripts in this paper (unless stated otherwise; e.g., V_i is a function, not a value in $\mathbf{R}$).

One of the important features of the above approach is that, even if the dimension of $\mathscr{H}$ is infinite, the solution belongs to a finite-dimensional subspace. In fact, under mild assumptions on the loss, according to the representer theorem [20], [16], $\hat{F}$ in (1) is the sum of kernel sections $K_i : \mathscr{X} \to \mathbf{R}$ defined by $K_i(x) = K(x_i, x)$. To be specific,

$$\hat{F}(\cdot) = \sum_{i=1}^{N} \hat{c}_i K_i(\cdot) , \qquad (2)$$

where $\hat{c}$ is defined by

$$\hat{c} = \arg\min_{c \in \mathbf{R}^N} \left(\sum_{i=1}^{N} V_i \left[y_i - \sum_{j=1}^{N} K(x_i, x_j) c_j \right] + \gamma c^{\mathrm{T}} \overline{K} c \right) . \qquad (3)$$

Here and below, $\overline{K} \in \mathbf{R}^{N \times N}$ is the so called *kernel matrix*, or Gram matrix, defined by $\overline{K}_{ij} = K(x_i, x_j)$. When the component loss functions $V_i(\cdot)$ are quadratic, the problem in (1) admits the structure of a regularization network [13] and also has a statistical interpretation: $\hat{F}$ is the minimum variance estimate of a Gaussian random field, given measurements corrupted by Gaussian white noise and a prior covariance proportional to K. This connection, briefly reviewed in Section III, is well known in the literature and was initially studied in [10] in the context of spline regression, see also [19], [7], [15]. It can be proved using the representer theorem together with the closed form solutions of the coefficients $\hat{c}_i$ in (2):

$$\hat{c} = \left(\overline{K} + \gamma \mathbf{I}_N \right)^{-1} y , \qquad (4)$$

where $y \in \mathbf{R}^N$ is the vector of measurements y_i and $\mathbf{I}_N$ is the $N \times N$ identity matrix.

A formal statistical model for more general loss functions (e.g., the Vapnik ε-insensitive loss used in support vector regression [18], [5], [8]) is missing from the literature. After interpreting the V_i as alternative statistical models for the observation noise, many papers argue that $\hat{F}$ in (1) can be viewed as a maximum a posteriori (MAP) estimator assuming the a priori probability density of F proportional to $\exp(-\|F\|^2_{\mathscr{H}})$, e.g. see Section 7 in [5]. These kinds of statements are informal, since in an infinite-dimensional function space the concept of probability density is not well defined, see e.g. [3] for a thorough treatment of Gaussian measures. The main contribution of this note is to provide a statistical model that justifies $\hat{F}$ as an estimate of a Gaussian random field.

The structure of the paper is as follows. In Section II we formulate the statistical model. In Section III, we review the connection between regularized estimation in RKHS and estimation in the quadratic case, and then extend this connection to more general losses. Section IV contains a summary and conclusion. The proofs are presented in Section V.

Aleksandr Y. Aravkin (saravkin@us.ibm.com) is with the IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

Bradley M. Bell (bradbell@uw.edu) is with the Applied Physics Laboratory & Institute for Health Metrics and Evaluation, University of Washington, Seattle WA, USA

James V. Burke (burke@math.washington.edu) is with the Department of Mathematics, University of Washington, Seattle, WA, USA

G. Pillonetto (giapi@dei.unipd.it) is with the Department of Information Engineering, University of Padova, Padova, Italy

This research has been partially supported by the European Community's Seventh Framework Programme [FP7/2007-2013] under agreement n. FP7-ICT-223866-FeedNetBack, under grant agreement n257462 HYCON2 Network of excellence and by the FIRB project entitled "Learning meets time".

II. Statistical Model

Here and below, $\mathbf{E}[\cdot]$ indicates the expectation operator and, given (column) random vectors u and v, and we define

$$\mathrm{cov}[u,v] = \mathbf{E}\left[(u-\mathbf{E}[u])(v-\mathbf{E}[v])^{\mathrm{T}}\right] .$$

We assume that the measurements y_i are obtained by sampling the function F in the presence of additive noise, i.e.

$$y_i = F(x_i) + e_i, \quad i = 1,\ldots,n , \tag{5}$$

where each x_i is a known sampling location. We make the following assumptions:

Assumption 1: We are given a known positive definite autocovariance function K on $\mathscr{X}\times\mathscr{X}$ and a scalar $\lambda > 0$ such that for any sequence of points $\{x_j : j = 1,\ldots,J\}$, the vector $f = [F(x_1),\ldots,F(x_J)]$ is a Gaussian random variable with mean zero and covariance given by

$$\mathrm{cov}(f_j, f_k) = \lambda K(x_j, x_k) .$$

■

A random function F that satisfies Assumption 1 is often referred to as a zero-mean Gaussian random field on $\mathscr{X}$.

Assumption 2: We are given a sequence of measurement pairs $(x_i, y_i) \in \mathscr{X}\times\mathbf{R}$ and corresponding loss functions V_i for $i = 1,\ldots,N$. In addition, we are given a scalar $\sigma > 0$ such that

$$\mathbf{p}(y|F) \propto \prod_{i=1}^{N} \exp\left(-\frac{V_i[y_i - F(x_i)]}{2\sigma^2}\right) .$$

Furthermove, the measurement noise random variables $e_i = y_i - F(x_i)$ are independent of the the random function F. ■

For example, $V_i(r) = r^2$ corresponds to Gaussian noise, while using $V_i(r) = |r|$ corresponds to Laplacian noise. The statistical interpretation of an ε-insensitive V_i in terms of Gaussians with mean and variance described by suitable random variables can be found in [14].

III. Maximum a posteriori estimation and reproducing kernel Hilbert spaces

A. Gaussian measurement noise

We first consider the case of Gaussian measurement noise; i.e., $V_i(r) = r^2$. This corresponds to modeling the $\{e_i\}$ as i.i.d. Gaussian random variables with variance σ^2. In view of the independence of F and e, it turns out that $F(x)$ and y are jointly Gaussian for any $x \in \mathscr{X}$. Hence, the posterior $\mathbf{p}[F(x)|y]$ is also Gaussian. The mean and variance for this posterior can be calculated using the following proposition [1, Example 3.6].

Proposition 3: Suppose u and v are jointly Gaussian random vectors. Then, $\mathbf{p}(u|v)$ is also Gaussian with mean and autocovariance given by

$$\begin{aligned} \mathbf{E}(u|v) &= \mathbf{E}(u) + \mathrm{cov}(u,v)\mathrm{cov}(v,v)^{-1}[v - \mathbf{E}(v)] , \\ \mathrm{cov}(u,u|v) &= \mathrm{cov}(u,u) - \mathrm{cov}(u,v)\mathrm{cov}(v,v)^{-1}\mathrm{cov}(v,u) . \end{aligned}$$

■

Suppose Assumptions 1, 2, and $V_i(r) = r^2$ for $i = 1,\ldots,N$. It follows that y is Gaussian. Applying Proposition 3 with $u = F(x)$ and $v = y$, we obtain $\mathbf{E}(u) = 0$, $\mathbf{E}(v) = 0$, and

$$\mathbf{E}[F(x)|y] = \lambda\begin{bmatrix} K_1(x) & \ldots & K_N(x)\end{bmatrix}\left(\lambda\overline{K} + \sigma^2\mathbf{I}_N\right)^{-1} y .$$

Using the notation $\gamma = \sigma^2/\lambda$, one obtains

$$\begin{aligned} \mathbf{E}[F(x)|y] &= \begin{bmatrix} K_1(x) & \ldots & K_N(x)\end{bmatrix}\left(\overline{K} + \gamma\mathbf{I}_N\right)^{-1} y , \\ &= \sum_{i=1}^{N} \hat{c}_i K_i(x) . \end{aligned}$$

where $\hat{c}$ is computed using (4). This expression shows that the minimum variance estimate coincides with $\hat{F}$ defined by (1). This result is formalized in the following proposition.

Proposition 4: Suppose that F satisfies Assumption 1 and $\mathbf{p}(y|F)$ satisfies Assumption 2 with $V_i(r) = r^2$. Then, the minimum variance estimate of $F(x)$ given y is $\hat{F}(x)$ defined by (1), with $\gamma = \sigma^2/\lambda$ and $\mathscr{H}$ the RKHS induced by K. ■

B. More general measurement noise

We now consider what happens when the Gaussian assumptions on e_i are removed. If the probability density function for F was well defined and given by

$$\mathbf{p}(F) \propto \exp\left(-\frac{\|F\|_{\mathscr{H}}^2}{2\lambda}\right) ,$$

then the posterior density conditional on the data would be

$$\mathbf{p}(F|y) \propto \exp\left(-\sum_{i=1}^{N}\frac{V_i[y_i - F(x_i)]}{2\sigma^2} - \frac{\|F\|_{\mathscr{H}}^2}{2\lambda}\right) .$$

In this case, the negative log of $\mathbf{p}(F|y)$ would be proportional to the objective in (1). Hence, one could immediately conclude that $\hat{F}$ is the MAP estimator. Unfortunately, the posterior density of F on a function space is not well defined, and this explains the somewhat indirect formulation of Proposition 5 below. Note that this proposition refers to the MAP estimator for f, a finitely sampled version of F.

Proposition 5: Suppose that F satisfies Assumption 1 and $\mathbf{p}(y|F)$ satisfies Assumption 2. Let M be any non-negative integer, $\{x_i : i = N+1,\ldots,N+M\}$ be an arbitrary set of points in $\mathscr{X}$, and define

$$f = [F(x_1),\ldots,F(x_{N+M})]^{\mathrm{T}} .$$

Then the MAP estimate for f given y is

$$\arg\max_f \mathbf{p}(y|f)\mathbf{p}(f) = [\hat{F}(x_1),\ldots,\hat{F}(x_{N+M})]^{\mathrm{T}} ,$$

where $\hat{F}$ is defined by (1), with $\gamma = \sigma^2/\lambda$ and $\mathscr{H}$ is the RKHS induced by K. ■

Remark 6: When considering non-Gaussian loss functions, the minimum variance and MAP estimates are different. Consider e.g. the case where $N = 1$, $M = 0$, $V_1(r) = |r|$, $y = 1$, and $\lambda = 1$, $\sigma = 1$, $K(x_1,x_1) = 1$. For this case, $f = F(x_1)$, and the MAP estimate for f given y is

$$\hat{f} = \arg\min_f(f^2 + |1-f|) = 1/2 . \tag{6}$$

Define A by

$$A = \int_{-\infty}^{+\infty} \exp(-f^2 - |1-f|)\mathbf{d}f .$$

The minimum variance estimate minus the MAP estimate is

$$\mathbf{E}(f|y) - \hat{f} = \frac{\exp(-3/4)}{A}\int_{1/2}^{+\infty} s\,\frac{\exp(1-2s)-1}{\exp(s^2)}\mathbf{d}s . \tag{7}$$

It follows that $A > 0$, for $s > 1/2$, the integrand above is negative, the righthand side is negative, and $\mathbf{E}(f|y) < \hat{f}$. ■

IV. Conclusion

When the RKHS induced by K is infinite-dimensional, the realizations of the Gaussian random field with autocovariance K do not fall in $\mathscr{H}$ with probability one, see eq. 34 in [12] and also [9], [4], [11] for generalizations. A simple heuristic argument illustrating this fact can be also found in Chapter 1 of [19]. The intuition here is that the realizations of F are much less regular than functions in the RKHS whose kernel is equal to the autocovariance K. On the other hand, in the case of Gaussian measurement noise, $\hat{F}$ defined in (1) is the minimum variance estimate; see Proposition 4. In this note we proved a formal connection between Bayesian estimation and the more general case prescribed by Assumption 2. For a given data set $\{(x_i, y_i)\}$, the RKHS function estimate $\hat{F}$ contains all the possible maximum a posteriori estimates of a random field for which the prior distribution is Gaussian. This result extends to more general cases by using more general versions of the Representer Theorem; see (2).

V. Appendix

A. Lemmas

We begin the appendix with two lemmas which are instrumental in proving Proposition 5:

Lemma 7: Suppose that g and h are jointly Gaussian random vectors. It follows that

$$\max_h \log \mathbf{p}(h|g) = -\log\det\left\{2\pi\left[\mathrm{cov}(h,h) - \mathrm{cov}(h,g)\mathrm{cov}(g,g)^{-1}\mathrm{cov}(g,h)\right]\right\}/2 ,$$

and this maximum does not depend on the value of g.

Proof: The proof comes from well known properties of joint Gaussian vectors, see e.g. [1]. The conditional density $\mathbf{p}(h|g)$ is Gaussian and is given by

$$\begin{aligned} -2\log\mathbf{p}(h|g) &= \log\det[2\pi\mathrm{cov}(h,h|g)] \\ &\quad + [h - \mathbf{E}(h|g)]^{\mathrm{T}}\mathrm{cov}(h,h|g)^{-1}[h - \mathbf{E}(h|g)] , \end{aligned}$$

where, recalling also Proposition 3,

$$\mathrm{cov}(h,h|g) = \mathrm{cov}(h,h) - \mathrm{cov}(h,g)\mathrm{cov}(g,g)^{-1}\mathrm{cov}(g,h) .$$

Thus, $\mathrm{cov}(h,h|g)$ does not depend on the value of g (and it would not make sense for it to depend on the value of h). Hence, one has

$$\begin{aligned} \arg\max_h \mathbf{p}(h|g) &= \mathbf{E}(h|g) , \\ \max_h \log\mathbf{p}(h|g) &= -\log\det[2\pi\mathrm{cov}(h,h|g)]/2 . \end{aligned}$$

This equation, and the representation for $\mathrm{cov}(h,h|g)$ above completes the proof of this lemma. ■

Lemma 8: Assume that g and h are jointly Gaussian random vectors and that y is a random vector such that $\mathbf{p}(y|g,h) = \mathbf{p}(y|g)$, and suppose we are given a value for y. Define the corresponding estimates for g and h by

$$(\hat{g},\hat{h}) = \arg\max_{g,h} \mathbf{p}(y,g,h) ,$$

and assume the above maximizers are unique. It follows that

$$\hat{g} = \arg\max_g \mathbf{p}(y|g)\mathbf{p}(g) , \tag{8}$$

$$\hat{h} = \arg\max_h \mathbf{p}(h|g=\hat{g}) . \tag{9}$$

Proof: We have

$$\begin{aligned} \mathbf{p}(y,g,h) &= \mathbf{p}(y|g,h)\,\mathbf{p}(h|g)\,\mathbf{p}(g) , \\ &= \mathbf{p}(y|g)\,\mathbf{p}(g)\,\mathbf{p}(h|g) , \\ \max_{g,h}\mathbf{p}(y,g,h) &= \max_g\left\{[\mathbf{p}(y|g)\,\mathbf{p}(g)]\,\max_h \mathbf{p}(h|g)\right\} . \end{aligned}$$

It follows from Lemma 7 that $\max_h \mathbf{p}(h|g)$ is constant with respect to g. Hence

$$\hat{g} = \arg\max_g [\mathbf{p}(y|g)\,\mathbf{p}(g)] ,$$

which completes the proof of (8). Hence

$$\max_{g,h}\mathbf{p}(y,g,h) = \mathbf{p}(y|\hat{g})\,\mathbf{p}(\hat{g})\max_h \mathbf{p}(h|g=\hat{g}) ,$$

which completes the proof of (9). ■

B. Proof of Proposition 5

Proof: The kernel matrix $\overline{K}$ is positive definite and hence invertible (Assumption 1). Define the random vectors g and h by

$$\begin{aligned} g &= [F(x_1),\ldots,F(x_N)]^{\mathrm{T}} , \\ h &= [F(x_{N+1}),\ldots,F(x_{N+M})]^{\mathrm{T}} . \end{aligned}$$

It follows that f in Proposition 5 is given by $f = (g^{\mathrm{T}}, h^{\mathrm{T}})^{\mathrm{T}}$. Notice that $\mathbf{p}(y|f) = \mathbf{p}(y|g)$ and that Lemma 8 can be applied. From (8) and the hypotheses above, we obtain

$$\begin{aligned} \hat{g} &= \arg\max_g \mathbf{p}(y|g)\mathbf{p}(g) , \\ &= \arg\max_g \left(\frac{1}{2\sigma^2}\sum_{i=1}^N V_i[y_i - g_i] + \frac{g^{\mathrm{T}}\overline{K}^{-1}g}{2\lambda}\right) , \end{aligned}$$

Using the representation $g = \overline{K}c$ we obtain

$$\hat{c} = \arg\max_c \left(\frac{1}{2\sigma^2}\sum_{i=1}^N V_i\left[y_i - \sum_{j=1}^N K(x_i,x_j)c_j\right] + \frac{c^{\mathrm{T}}\overline{K}c}{2\lambda}\right) .$$

This agrees with (3), because $\gamma = \sigma^2/\lambda$, and thereby shows

$$\hat{g} = [\hat{F}(x_1),\cdots,\hat{F}(x_N)]^{\mathrm{T}} .$$

Finally, by Proposition 3 and Lemma 7 in conjuction with (2), (9), and the expression for $\hat{g}$ above, we obtain

$$\begin{aligned} \hat{h} &= \mathrm{cov}(h,g)\mathrm{cov}(g,g)^{-1}\hat{g} \\ &= \mathrm{cov}(h,g)(\lambda\overline{K})^{-1}(\overline{K}\hat{c}) , \\ &= \begin{pmatrix} K_1(x_{N+1}) & \ldots & K_N(x_{N+1}) \\ \vdots & \ddots & \vdots \\ K_1(x_{N+M}) & \ldots & K_N(x_{N+M}) \end{pmatrix}\begin{pmatrix} \hat{c}_1 \\ \vdots \\ \hat{c}_N \end{pmatrix} , \\ &= [\hat{F}(x_{N+1}),\ldots,\hat{F}(x_{N+M})]^{\mathrm{T}} . \end{aligned}$$

Combining this with the formula for $\hat{g}$ above, we conclude

$$[\hat{F}(x_1),\ldots,\hat{F}(x_{N+M})]^{\mathrm{T}} = \arg\max_f \mathbf{p}(y,f) ,$$

which completes the proof of Proposition 5. ■

C. Proof of Remark 7

Proof: It suffices to show that equations (6) and (7) hold. It follows from $N=1$, $\gamma=1$, that c is a scalar, $f=F(x_1)=c$, and using (3) we have

$$\hat{f}=\hat{c}=\arg\min_c |1-c|+c^2=1/2$$

It also follows that

$$\mathbf{p}(y|f)\mathbf{p}(f)\propto \exp(-f^2-|1-f|)$$

The minimum variance estimate $\mathbf{E}(f|y)$, and its difference from the map estimate $\hat{f}$, are given by

$$\begin{aligned}\mathbf{E}(f|y)&=\frac{1}{A}\int_{-\infty}^{+\infty} f\exp(-f^2-|1-f|)\mathbf{d}f\,,\\ \mathbf{E}(f|y)-\hat{f}&=\frac{1}{A}\int_{-\infty}^{1}(f-1/2)\exp(-f^2-1+f)\mathbf{d}f\\ &\quad+\frac{1}{A}\int_{1}^{+\infty}(f-1/2)\exp(-f^2+1-f)\mathbf{d}f\,.\end{aligned}$$

Multiplying both sides of the equation by A and using the change of variables $s=f-1/2$, we obtain

$$\begin{aligned}&A(\mathbf{E}(f|y)-\hat{f})\\ &=\int_{-\infty}^{1/2} s\exp[-(s+1/2)^2+s-1/2]\mathbf{d}s\\ &\qquad+\int_{1/2}^{+\infty} s\exp[-(s+1/2)^2-s+1/2]\mathbf{d}s\,,\\ &=\int_{-\infty}^{1/2} s\exp(-s^2-3/4)\mathbf{d}s\\ &\qquad+\int_{1/2}^{+\infty} s\exp(-s^2-2s+1/4)\mathbf{d}s\,,\\ &=\int_{-\infty}^{-1/2} s\exp(-s^2-3/4)\mathbf{d}s\\ &\qquad+\int_{1/2}^{+\infty} s\exp(-s^2-2s+1/4)\mathbf{d}s\,,\\ &=\int_{1/2}^{+\infty} s\exp(-s^2-3/4)[\exp(1-2s)-1]\mathbf{d}s\,.\end{aligned}$$

This completes the proof of the remark. ■

REFERENCES

[1] B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, N.J., USA, 1979.

[2] N. Aronszajn. Theory of reproducing kernels. *Trans. of the American Mathematical Society*, 68:337–404, 1950.

[3] V.J. Bogachev. *Gaussian measures*. AMS, 1998.

[4] M. Driscoll. The reproducing kernel Hilbert space structure of the sample paths of a Gaussian process. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 26:309–316, 1973.

[5] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–150, 2000.

[6] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architecture. *Neural Computation*, 7:219–269, 1995.

[7] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, 1995.

[8] L. Gunter and J. Zhu. Computing the solution path for the regularized support vector regression. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 483–490. MIT Press, Cambridge, MA, 2006.

[9] G. Kallianpur. The role of reproducing kernel Hilbert spaces in the study of Gaussian processes. In *Advances in Probability and Related Topics*, pages 49–83. Marcel Dekker, 1970.

[10] G. Kimeldorf and G. Wahba. A correspondence between Bayesan estimation of stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41(2):495–502, 1971.

[11] M.N. Lukic and J.H. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Trans. Amer. Math. Soc.*, 353:3945–3969, 2001.

[12] E. Parzen. Probability density functionals and reproducing kernel Hilbert spaces. In *Proc. of the Symposium on Time Series Analysis*, New York, 1963. John Wiley and Sons.

[13] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78:1481–1497, 1990.

[14] M. Pontil, S. Mukherjee, and F. Girosi. On the noise model of support vector machine regression. In *Proc. of Algorithmic Learning Theory 11th International Conference ALT 2000*, Sydney, 2000.

[15] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[16] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. *Neural Networks and Computational Learning Theory*, 81:416–426, 2001.

[17] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. (Adaptive Computation and Machine Learning). The MIT Press, 2001.

[18] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, USA, 1998.

[19] G. Wahba. *Spline models for observational data*. SIAM, Philadelphia, 1990.

[20] G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and randomized GACV. Technical Report 984, Department of Statistics, University of Wisconsin, 1998.